Whitepaper

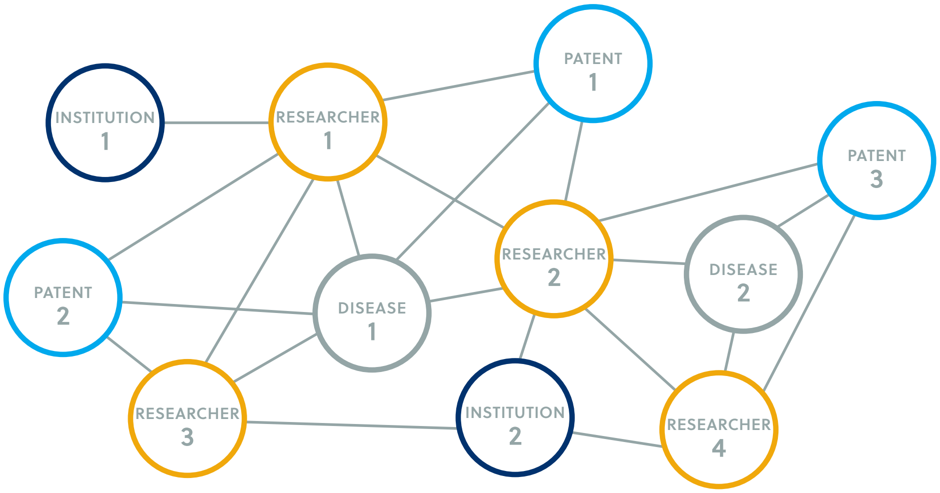
Knowledge Graphs

Connecting Your Data to Solve Real-World Problems in R&D, Business Intelligence, and Strategy

Phill Jones, PhD
Double L Digital

The volume and types of information and knowledge that R&D intensive companies must ingest, process, and synthesize is increasing super-exponentially. In the past, this has forced knowledge and information professionals to effectively silo information by subscribing to databases and content sources that contain specific types of objects or relate to certain subject areas. In response, many knowledge and information professionals feel that something was lost along the way; the serendipity of finding a piece of information in a place you did not realize you needed to look[1].

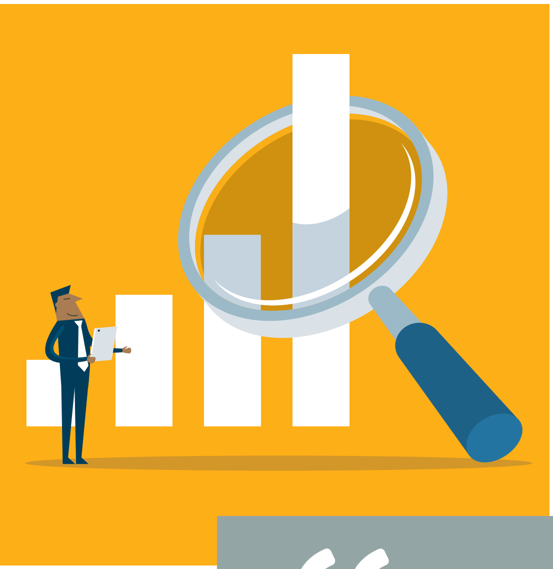
What is a Knowledge Graph?



A knowledge graph is a collection of information objects, which could be data, documents, and/or metadata about people, places, institutions, or anything else. In graph theory those objects are sometimes called ‘nodes.’ The nodes are connected to each other by links that represent relationships. Those links can somewhat counter-intuitively be referred to as ‘edges’ of the graph.

In this paper, we will explore the development and application of knowledge graphs and how they help knowledge and information professionals solve real business problems. Knowledge graphs treat the connections between pieces of information with as much importance as the information itself. Exposing these connections across data sources of all shapes and sizes allows end users to ask meaningful, business relevant questions and get actionable answers.

We will also share some specific examples of business problems for which knowledge graphs are already being used, based on a series of conversations with forward-thinking professionals in the information science and knowledge management space. Finally, we will present a series of tips to help you develop your own knowledge graph roadmap. This includes advice on how to clean and map data and create a data processing pipeline to position your organization to benefit from the greatest insight from both its data and the connections between them.



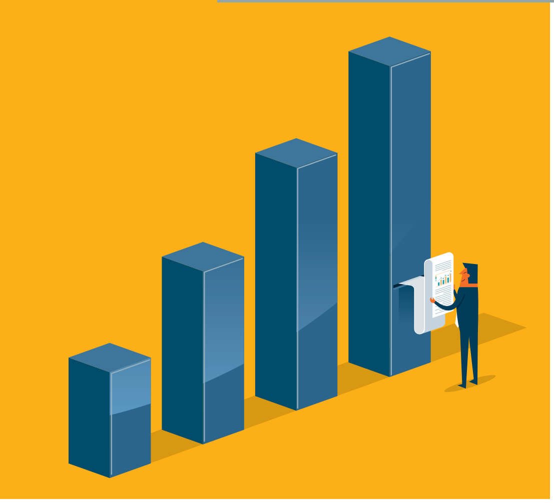
Finding answers to questions by reviving The art of serendipitous discovery

Professionals in knowledge and information management are looking for inventive ways to enable the researchers, analysts and executives they support to find relevant information faster and make more informed decisions. The goal is to make discoveries and find connections irrespective of the origin of that information. As one senior information executive at a major pharmaceutical company said:

> “…we don’t even provide a library catalog anymore, we just do it through search. Essentially, we’re saying, you don’t need to worry about where the information or the data is coming from. Just look for the data that you’re after.”

This is a journey information and knowledge managers have been on for some time as the state of the art has evolved from separate indexing services and databases, to federated and indexed search, and eventually fully aggregated search that combines the best of all worlds[3]. The creation of knowledge graphs promises to further invigorate the discovery process by offering connected insights researchers did not realize they needed. While this is entirely by design and certainly no accident, the ability to find useful related information in a way that feels natural and intuitive lies at the heart of what some people call “serendipitous discovery,” and is a key way in which knowledge graphs add value. As it was described to me by a director of information and intelligence at a pharmaceutical company:

> We've moved on from federated search, if you like, it's now about intelligent search. That predictive search that says, 'right, this is what I'm interested in'. Now, I need to make the connections with that topic or object that I'm interested in. And then I might also want to go deeper to make connections between data points that I had no insight over previously."



You are already an expert user of knowledge graphs

The concept of a knowledge graph is deceptively simple and almost everybody is familiar with it, even if many end users do not recognize the name. That is because anybody working in the service or knowledge economies likely makes use of the best-known example of one every single day: the world wide web. The web is mostly a series of documents connected to each other through hyperlinks that enable a reader to discover information by traversing those links as they come across ideas or content that they find interesting. Increasingly, of course, web browsing habits are being driven by search, as the technologies to index, capture and efficiently identify and extract the information the user is likely to be looking for continue to become more powerful.

Modern database design is about information use, not just data storage

Since the computing revolution of the mid-20th century, there have been a number of evolutions in how we store data and information[5]. The earliest database used hierarchical structures that were heavily constrained by storage space and computing power. Over time, more efficient structures like the Entity-Relationship Model[6] were developed that use tables, rather like a spreadsheet. The best-known example of this type is SQL.

While efficient, tables do not lend themselves easily to use cases that involve unstructured data like documents and are not designed with the needs of the information user in mind. The rise of web-based applications and standards[7], coupled with increases in computing power, led to the era of NoSQL (Not Only SQL) in database design. Data scientists were able to think beyond the most efficient storage structures and put more emphasis on what end users need from the information. This led to more specialist database types like document databases and knowledge graphs[8].

How knowledge graphs answer scientific questions

Knowledge graphs are powerful because they shift the focus toward applying information to real questions and answers. This is achieved by treating the connections between the objects as important as the objects themselves. Connections between entities, topics, indications, and assertions, make it possible to ask questions and get meaningful answers. Another way of putting it: knowledge graphs preserve the complexity inherent in information, enabling complex problems to be better addressed.

A head of information solutions at one of the world's largest pharmaceutical companies summed it up as follows:

> "The true value of knowledge graphs is that they allow us to identify relationships that exist between documents, not just within a single document... classical searches that use Boolean techniques can return multiple documents that contain sort of what you're looking for... where you've got a knowledge graph you can put in, say three terms and ask what could connect term A with term B or term C. A lot of the time it's doing what a scientist naturally does in their head."

The Persistent Identifier (PID) graph

One way the global information landscape is becoming more robust is through persistent identifiers. According to a recent report from the national research organization Jisc in the UK[9], there are a series of five priority PIDs needed to interconnect people, organizations, research grants, projects, and outputs like articles and data sets.

A recent post on the project Freya blog[10] concisely describes how the linking of PIDs to others through appropriate metadata can drastically improve discovery of, for example, all the data sets associated with a project, or the researchers that have worked on a project or at an institution.

Here is an example. Imagine a research team is working on a new drug that might be effective against a particular disease. A simple literature search might find multiple instances of the indication or even the biological target mentioned alongside a particular molecule. However, we often need to go deeper than that and explore potential side effects, predict what the best mechanism for administration might be, or find ways to avoid issues through medicinal chemistry and molecule design. This is where a knowledge graph can help researchers make those leaps of understanding by connecting apparently disparate pieces of information. The same knowledge and information systems expert said:

> “They [scientists] will connect terms. They will say ’oh, that gene, I’m sure I’ve seen that linked to this question somewhere’. [The] power of a knowledge graph is you don’t need to have that in-depth knowledge. It can quickly pull out 10 articles. Five of them say this drug hits that target and five of them show that target is associated with a physical symptom.”

Other scientific areas where graphs can help include gene-disease associations which are easy for people to miss due to the sheer volume of information and inconsistencies in taxonomies across fields, and linking causal effects associated with diseases for early target identification.

Graphs can solve a diverse set of business problems

The use of graph technology is not limited to scientific questions. Knowledge and information managers are curating and processing a diverse range of sources that are useful for multiple functions across an organization, from business intelligence and strategy to legal affairs and financial planning. So how can these functions be better served by knowledge graphs? The question to ask as a knowledge and information manager is ‘where do the connections between those sources add value?’

Empowering business intelligence

A group working in business intelligence, for example, might be interested in which drugs are being developed to treat a particular disease. This could require linking across scientific and medical literature as well as clinical trials databases. Going further, a full business intelligence landscape might draw from many more sources. Drug pipeline and clinical trials data can tell a team much about what competitors are investing in. Finding links between those pipelines and mergers and acquisitions from SEC filings, industry news, and patent filings can give further clues. It is not hard to see how a complex graph of disparate information sources with information coming in multiple shapes and sizes can be cross-referenced to gain extraordinary insight.

Reducing costs in clinical trial design

One of the experts with whom I spoke was a director at a company that provides clinical development analysis products. They use data from clinical trials databases, as well as some surprising sources like researcher CVs and financial reports, to improve clinical trial design and investment decisions. Some of the questions that their group was able to answer included how to optimize trial design based on number and locations of sites. They also help triage trials during difficult financial times to maximize return on investment and reduce risks. The director said:

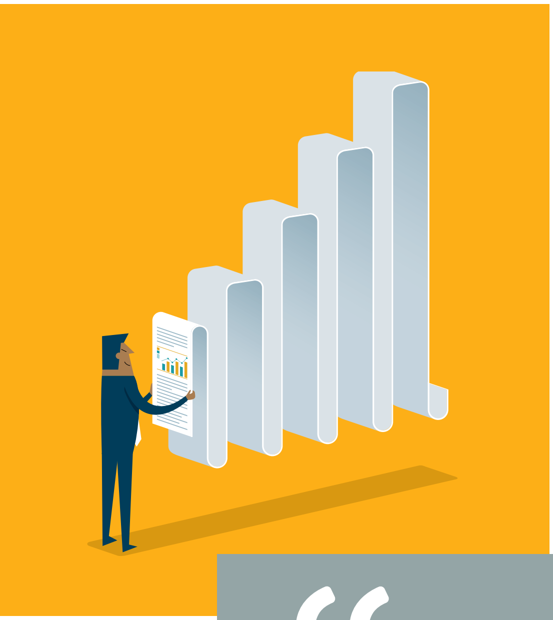
> “Even before COVID, 20% of those trials would have failed. Now, I don't know what the number is, but it's probably a lot worse. A lot of those trials are in trouble and the companies that do those trials have challenges as a result. It's possible to analyse the data and triage those that are most likely to work out poorly and improve the effective hit rate, thereby saving significant costs.”

As well as saving costs, finding previously missed connections can also improve the financial viability of a project. In their recent project, they were able to expand the indication of the trial drug using knowledge graphs and thereby rescue the at-risk project, according to the person I interviewed.

> “We were recently working on a trial where the therapeutic area was iron overload. After using our tools to look at the therapeutic agent, we were able to expand the indication in a way that attracted ten times as much funding.”

Identifying the key opinion leaders faster

Another use case that an interviewee spoke of was identifying key leaders in particular fields of research. This might be of use for R&D when looking for academic collaborators. It might also be of use for HR departments engaged in talent acquisition. Knowledge graphs can add particular value in areas that are emerging or cross-disciplinary, where it is not always clear who the new and emerging leaders are. Knowledge graphs can spot emergent connections in new areas even before communities recognize them.



COVID-19 is an emergent trans-disciplinary field

A very pertinent and recent example of the need to identify leaders quickly is in the fight against COVID-19[11]. In the bottom-right corner of Figure 1, the bar chart shows the number of articles indexed per month in *PubMed* returned from a Boolean search for *coronavirus*, covid, sars-cov, or sars-cov2. There has clearly been an explosion of activity in the last few months. In essence, a new trans-disciplinary research field has emerged in a period of months with researchers across the world applying for grants, repurposing effort, and looking for solutions.

With this rate of work, it would be impossible for a researcher to digest all of this content, find the connections, the most likely best approaches, and identify the leaders. The middle and top most panels, show how mapping the topic areas and authors within this new discipline can give rapid insight to this new and vital field. To quote a pharmaceutical industry expert in knowledge graph applications to whom I spoke:

> “The COVID example is good in that no single source can give you a strong picture of experts and activities... Classical document search tools will show you the number of publications, clinical trials, news articles, patents or grants.
>
> A graph database would connect all this together, so you would get a clearer picture if there were centers of activities or key drivers. You would also be able to see the historical expertise in the related areas that make the person’s activities relevant to COVID.”

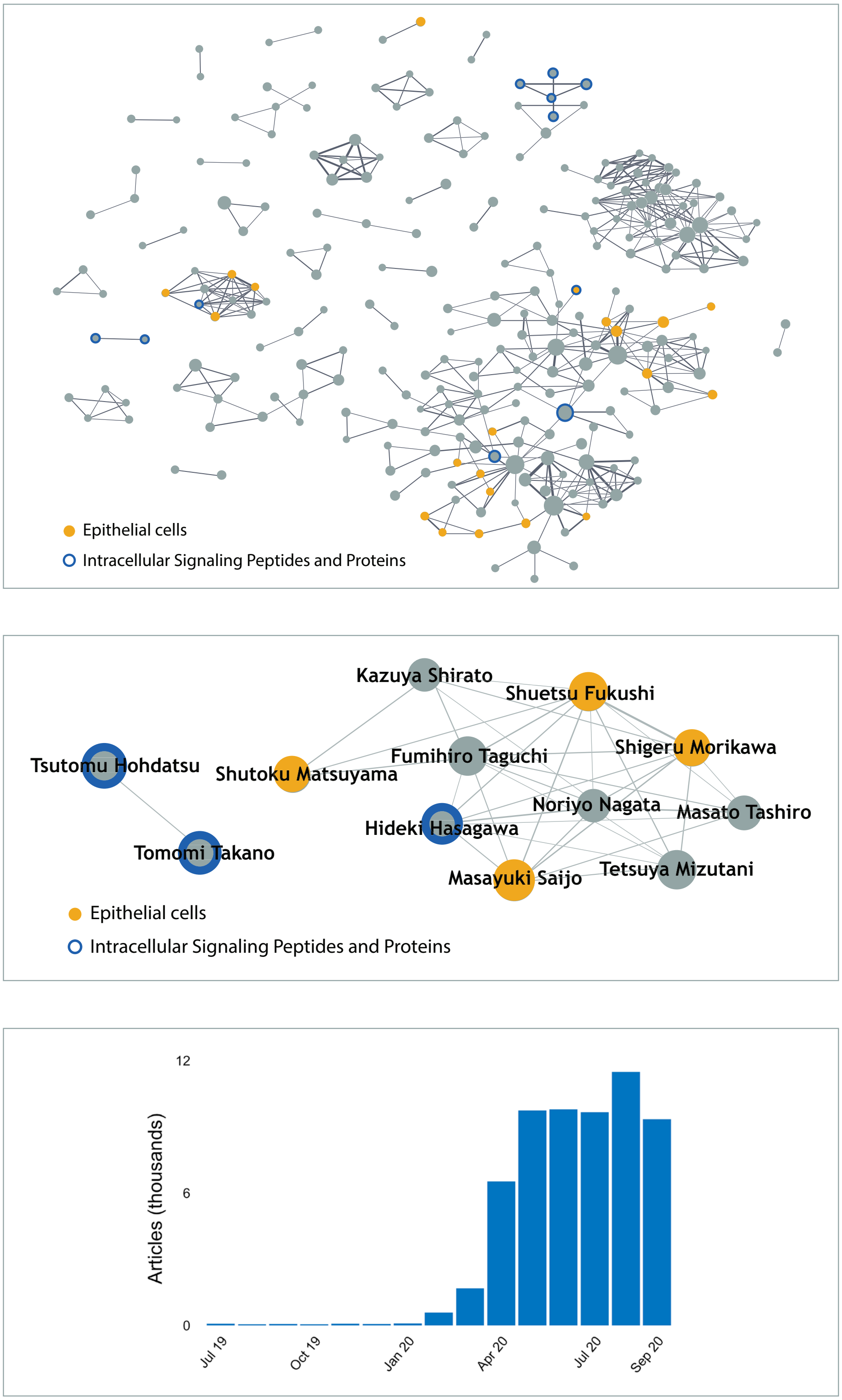


Figure 1: The bottom-most panel shows the number of research articles in PubMed about coronavirus or COVID-19 published per month between January 2019 and August 2020. The shear velocity of content in the new discipline would make it impossible for a human to keep up without the use of a knowledge graph to find the connections between the data. The top panel shows a knowledge graph where the nodes are authors based on the CORD-19 and LitCovid datasets. The sizes of the nodes indicate how many articles they've published relating to COVID-19 and the thickness of the edges connecting the nodes indicate the number of collaborations. We highlight two sub-networks using MeSH keywords, one for researchers with expertise in epithelial cells, the other shows researchers with expertise in intracellular signaling peptides and proteins. Visual exploration helps us to quickly generate hypotheses. For example, a local network shown on the right of the middle panel contains several experts in epithelial cells, but only one in intracellular signaling. An adjacent network of just two researchers have expertise in intracellular signaling but not in epithelial cells. Perhaps a collaboration would be mutually beneficial!

Tips on how to build and work with knowledge graphs

The interviewees I spoke with for this paper were clear in pointing out that not all knowledge graphs are created equal and some approaches yield better results than others. Here is some of the advice they gave.

Do not throw away information unnecessarily

This first piece of advice speaks to the philosophy of knowledge graphs. Traditional information management approaches often rely on forcing information into controlled structures that inherently impose assumptions.

A simple example is controlled vocabularies. If two words are decided to be a synonym, and therefore one is changed to the other in your database, any potential difference in meaning is lost. It may be better to map the relationship between those two terms, so that the relationship between them can be modified at a later date. To quote the head of a knowledge center a pharmaceutical company:

> “Make sure the data in the database is clean. And if it's not, then decide how you're going to clean it. Are you going to upgrade the databases and put in the controlled vocabularies? Or is that going to be the first part of your knowledge graph?”

Take the time to understand the data and information you have

While it is often better to not discard information, sometimes lack of data cleanliness creates the appearance of more information than there really is. It is therefore important to understand your data before attempting to put it into a knowledge graph. Make informed decisions about where data needs to be cleaned and where it needs to be mapped.

Define data governance metrics such as completeness and data quality at each stage of the data processing pipeline and conduct quality control to catch bad data. That can be as simple as detecting entries of a single character length, impossible entries like dates in the future, character set encoding errors and even gene names that have been coerced into dates[12].

Think about the questions your organization needs to ask

Each organization's data is unique and has its own management challenges driven, in part, by the types of business problems the organization needs to solve. Identification and deduplication of authors, for example, is a common challenge. Names are not unique, sometimes change, and can be styled in multiple ways. Mapping across affiliations and topics enables clusters of articles that likely share an author to be linked. Connections can be changed to refine the graph as new information becomes available. As one interviewee said:

> [you might] say you're going to use what's in the dictionaries and do the cleaning outside. You then just end up with all your mapping within your knowledge graph. [You can] decide to what data you want to connect and what is the best way to connect it..."

Remember the value of knowledge graphs lies in the connections between objects. Those connections are answers to questions and surface the value of the data and the information. It is therefore important to design the connections to answer the sorts of questions you need to ask, which may be similar to the questions suggested in this white paper, or may be very different.

Maintain and show provenance

One criticism of Google's knowledge graph is that it does not tell the users enough about the reliability of its sources[13]. If the interface you use does not show researchers and executives where information comes from, they are unlikely to believe it and might be vocal in expressing such. One interviewee noted:

> ...it shouldn't be based on the person who's got the loudest voice, who may have just pulled the data from [source] and made a typo when they put it in. You have to keep the provenance. If somebody disagrees with the capture, they need to be able to say why. Otherwise they may [just] overwrite that value. Then nobody knows what the correct answer is or why it seems to have changed."

Provide flexibility in how users interact with knowledge graph

Remember one of the big advantages of knowledge graphs is they are designed with the uses of the information in mind. The same knowledge graph may serve a range of user groups including researchers, analysts, executives, and even HR, but if they ask different questions, they may need the information presented in different ways.

For example, some knowledge graph interfaces allow users to visually navigate the graph, by jumping from node to node. This approach might be best used by data scientists or biomedical researchers who have the domain expertise to interpret the information in a direct way.

For other users, having a knowledge graph interface could be too complicated and time consuming for discovering useful information or insights. For those users, perhaps in business intelligence or strategy, applying the knowledge graph under-the-hood, and simply presenting its results the way that a Google search does is likely preferable.

Let knowledge graphs change; they should be living objects

Thinking carefully about the design of the graph and how things are mapped is vital, but it is naive to think it will be perfect from the start. Allowing users to provide feedback is an excellent way to evolve the knowledge graph and make it even more powerful.

That feedback could be in the form of requests to make different connections or ask different questions about the data. On the other hand, that feedback could be in the form of weighting of data points. If you maintain and show provenance, end users might, for example indicate a particular article has erroneous results or maybe a market sizing estimate is known to be wrong. Allowing users to feed back those judgements allows the capturing of that organizational tacit knowledge.

Create your own knowledge graph roadmap

As technology evolves and the amount of information available continues to grow, knowledge graphs will become increasingly vital as a tool for information and knowledge managers as well as the researchers and executives they support. Use the insights above to lay out your own knowledge graph roadmap. Find the connections in your data you never knew existed, solve real business problems and make exiting new discoveries.

Phill Jones, PhD

Double L Digital

Phill Jones has a decade of experience bringing innovative products to market. Prior to founding DLD, Phill was the CTO at Emerald Publishing. He has had a series of roles at Digital Science (DS), including a senior role in the DS Consultancy. He also led thought leadership efforts in scholarly publishing, and developed patron driven acquisition and article syndication business models. Phill was the first Editorial Director at Journal of Visualized Experiments and is an influential thought leader in the scholarly communications technology sector. Areas of expertise include product and technology strategy, market-led digital innovation, and the changing landscape of academia. Phill is a former cross-disciplinary researcher. He received a PhD in physics from Imperial College, London and held a faculty position in neuroscience at Harvard Medical School.

References

[1] L. Y. Conrad and P. D. Moeller, 'Search, Serendipity, and the Researcher Experience', *Ser. Libr.*, vol. 72, no. 1–4, pp. 190–193, May 2017, doi: 10.1080/0361526X.2017.1292744.

[2] R. Shields, 'Cultural Topology: The Seven Bridges of Königsburg, 1736':, *Theory Cult. Soc.*, Oct. 2012, doi: 10.1177/0263276412451161.

[3] R. Gilmartin, 'How Does Aggregated Search Work?', *Velocity of Content*, Jul. 07, 2020. https://www.copyright.com/blog/how-does-aggregated-search-work/ (accessed Sep. 12, 2020).

[4] 'Introducing the Knowledge Graph: things, not strings', *Official Google Blog*, 2012. https://web.archive.org/web/20180711042907/https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html?m=1 (accessed Sep. 10, 2020).

[5] M. Broecheler, *Practitioner's Guide to Graph Data: Applying Graph Thinking and Graph Technologies to Solve Complex Problems*. O'Reilly Media, Incorporated, 2020.

[6] P. P.-S. Chen, 'The entity-relationship model—toward a unified view of data', *ACM Trans. Database Syst.*, vol. 1, no. 1, pp. 9–36, Mar. 1976, doi: 10.1145/320434.320440.

[7] C. Mohan, 'History repeats itself: sensible and NonsenSQL aspects of the NoSQL hoopla', in Proceedings of the 16th International Conference on Extending Database Technology - EDBT '13, Genoa, Italy, 2013, pp. 11–16, doi: 10.1145/2452376.2452378.

[8] M. Rouse, 'What is NoSQL (Not Only SQL database)? - Definition from WhatIs.com', *SearchDataManagement*, 2017. https://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL (accessed Sep. 11, 2020).

[9] J. Brown, 'Developing a persistent identifier roadmap for open access to UK research', Jul. 2019. [Online]. Available: http://repository.jisc.ac.uk/id/eprint/7840.

[10] M. Fenner and Aryani, Amir, 'Introducing the PID Graph', *Freya Blog*, Mar. 28, 2019. https://web.archive.org/web/20190712173134/https://www.project-freya.eu/en/blogs/blogs/the-pid-graph (accessed Jul. 12, 2019).

[11] E. A. Holmes et al., 'Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science', *Lancet Psychiatry*, vol. 7, no. 6, pp. 547–560, Jun. 2020, doi: 10.1016/S2215-0366(20)30168-1.

[12] M. Ziemann, Y. Eren, and A. El-Osta, 'Gene name errors are widespread in the scientific literature', *Genome Biol.*, vol. 17, no. 1, p. 177, Aug. 2016, doi: 10.1186/s13059-016-1044-7.

[13] C. Dewey, 'You probably haven't even noticed Google's sketchy quest to control the world's knowledge', *Washington Post*.